


- 
- (i) parsing the genomics text data to determine the grammatical structure of the text data ;
  - (ii) regularizing the parsed text data to form structured word terms;  
and
  - (iii) tagging the text data with a structured data component derived from the structured word terms wherein said tagging step comprises providing the structured data component in a Standard Generalized Markup Language (SGML) compatible format.
- 

#### R E M A R K S

Claims 11-21 are pending in the application. Claims 11-21 are rejected under 35 U.S.C. §112, second paragraph; claims 11, 13-21 are rejected under 35 U.S.C. §§ 101, 102(e), 102(f) and claims 11-21 are rejected under 35 U.S.C. § 103(a). The claims have been amended to more particularly point out and distinctly claim the invention. No new matter is added. For reasons detailed below, the rejections should be withdrawn and the claims allowed to issue. Entry of the foregoing amendments is respectfully requested.

1. Objections to the Specification and Claims

The Examiner has objected to the specification because the disclosure contains embedded hyperlinks and/or other forms or browser executable codes. As requested by the Examiner, Applicants have deleted the embedded hyperlinks.

According to the Examiner, the title of the invention is not descriptive.

Applicants have amended the title to read as follows "Methods for Extracting Information on Interactions Between Biological Entities from Natural-Language Test Data."

In response to the request to submit drawing corrections, Applicants submit herewith a set of correct drawings.

Applicants have amended the first sentence of the specification to read as follows:

"This application is a continuation-in-part of pending application Serial No. 09/327,938 filed June 8, 1999 which claims priority to provisional patent application Serial No. 60/129,469 filed on April 15, 1999."

## 2. The Rejections Under 35 U.S.C. § 101 Should Be Withdrawn

Claims 11, 13-21 are rejected under 35 U.S.C. § 101 as claiming the same invention as that of claims 1-3, 5, 7, and 9-16 of prior U.S. Patent No. 6,182,029. The Examiner alleges that claims 11, 13-21 of the instant application and claims 1-3, 5, 7, and 9-16 of prior U.S. Patent No. 6,182,029 are directed to the same invention: method for extracting information from natural-language text data including steps in the same scope. The Examiner has noted that while the preambles of claims 11, 13-21 of the instant application recite "extracting information on interactions between biological entities from natural-language text data," yet the actual method steps do not involve at all "biological entities."

Applicants have amended claim 11 to specify that the parsing step relates to parsing of "*genomics*" text data. Support for amended claim 11 can be found on page 17, lines 2-5 of the specification. In view of the amendment to claim 11 which clearly

distinguishes the presently claimed invention from the claims of U.S. Patent No. 6,182,029, Applicants respectfully request that the rejection under 35 U.S.C. §101 be withdrawn.

3. The Rejections Under 35 U.S.C. § 112, Second  
Paragraph Should Be Withdrawn

Claims 11-21 are rejected under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which Applicants regard as the invention.

The Examiner alleges that the preambles of claim 11 and its dependent claims 13-21 recite "a method for extracting information on interactions between biological entities" but the method steps do not accomplish such and do not involve biological entities. Applicants have amended claim 11 to specify that "*genomics*" text data is parsed. Such genomics test data provides information relating to interactions among genes and proteins (see p. 16, lines 4-9 of the specification).

The Examiner maintains that the phrases "regularizing the parsed text data" and "natural language" in claim 11 and all its dependent claims is vague and indefinite. Further, according to the Examiner the phrase "undefined words" in claim 17 is vague and indefinite. The Examiner alleges that the phrase "binary actions" in claim 18 and the phrase "when parsing of the text data is unsuccessful" in claim 19 are also vague and indefinite.

Applicants assert that one skilled in the art would understand the meaning of the terms and phrases used in the claims. In this regard, the Examiner's attention is directed

to U.S. Patent 6, 182, 029 which was incorporated in its entirety by reference into the present specification (see p.4, lines 7-13). Specifically, col. 1, lines 37-43 define "natural language"; col. 6, line 65 through col. 7 line 7 defines "regularized"; col. 9, lines 43-50 defines "when parsing of text data is unsuccessful; and col. 10, lines 22-27 defines "undefined words."

In claim 19, the phrase "parsing of the text data is unsuccessful" is alleged by the Examiner to lack clear antecedent basis. Applicants have amended claim 19 to correct the antecedent basis.

In claim 21, the phrase "said tagging step" is said by the Examiner to lack antecedent basis. Applicants have amended claim 21 to correct the antecedent basis.

In view of the above, Applicants request withdrawal of the rejections under 35 U.S.C. § 112.

#### 4. The Claims Are Not Anticipated

Claims 11, 13-21 are rejected under 35 U.S.C. § 102(e) as being anticipated by Friedman C. (U.S. Patent No. 6,182,029; Date of Patent: January 30, 2001; Filed: August 6, 1999; "the '029 patent"). According to the Examiner, claims 11, 13-21 of the instant application and claims 1-3, 5, 7, and 9-16 of Friedman are directed to the same invention, i.e., method for extracting information from natural-language text data including steps in the same scope. Although the preambles of claims 11, 13-21 of the instant application recite "extracting information on interactions between biological entities from natural-language text data," the actual method steps do not involve at all "biological entities."

The Examiner maintains that claims 11, 13-21 are rejected under 35 U.S.C. § 102(f) because the Applicant did not invent the claimed subject matter. The Examiner asserts that Friedman C. (U.S. Patent No. 6,182,029; Date of Patent: January 30, 2001; Filed: August 6, 1999; "the '029 patent") invented the method for extracting information from natural-language text data including steps in the same scope as the steps of the method in claims 11, 13-21 of the instant application. Although the preambles of claims 11, 13-21 of the instant application recite "extracting information on interactions between biological entities from natural-language text data," the actual method steps do not involve at all "biological entities." Thus, the limitation bears no weight in the process of examination.

Applicants have amended claim 17 to specify that the method for extracting information on interactions between biological entities comprises, as a first step, parsing "*genomics*" text data. Given the distinction between the presently claimed invention and that of the '029 patent, Applicants respectfully request that the rejections under 35 U.S.C. § 102 (e) and (f) be withdrawn.

#### 5. The Claims Are Not Obvious

Claims 11-21 are rejected under 35 U.S.C. § 103(a) as being unpatentable over Friedman C. (U.S. Patent No. 6,182,029; "the '029 patent"). According to the Examiner, Friedman discloses methods for extracting information from natural-language text data comprising parsing the text data, preprocessing the data prior to parsing and regularizing the parsed data (see columns 5, and 16-17). The Examiner asserts that Friedman also suggests/motivates application of the extraction method in "extracting medical/clinical

data from physician reports and genomics-related information from electronic text records" (column 4). The Examiner alleges that it would have been well-known that genomics-related information is biological information. Further, Friedman actually uses such entities in a text as "proteins," "genes" and "activate" as examples to demonstrate how parsing works (column 6), and it would have been well-known that these entities are biological entities, as required in the instant claim 12. Thus, according to the Examiner, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to combine the teachings and/or suggestions/motivations of Friedman and what would have been well-known to make and use the invention because identifying biological entities such as "proteins," "genes" and "activate" in parsing is actually taught or suggested by Friedman.

Claims 11-21 are rejected under 35 U.S.C. § 103(a) as being unpatentable over Kim, K.H. (Comparative Molecular Field Analysis (CoMFA), 1995). The Examiner asserts that Kim reviews the state of the art in the field of comparative molecular field analysis (CoMFA) combining a great number of references, i.e., text data, and involving many biological entities, e.g., receptor molecule (see page 29, second and third paragraphs, and pages 323-324). According to the Examiner, it would have been obvious to one of ordinary skill in the art that in the process of preparing for the review article, Kim must have parsed each reference, i.e., read the reference article word by word, phrase by phrase, sentence by sentence and paragraph by paragraph and eventually regularized the parsed text data, i.e., formed the final version of the review article.

A finding of obviousness under §103 requires a determination of the scope and content of the prior art, the level of ordinary skill in the art, the differences between the

claimed subject matter and the prior art, and whether the differences are such that the subject matter as a whole would have been obvious to one of ordinary skill in the art at the time the invention was made. *Graham v. Deere*, 383 U.S. 1.

Applicants assert that the disclosure of the '029 patent relates to a natural language processing system for *extracting medical/clinical data from physician reports*. As such, terms in a natural language phrase relating to, for example, body parts and clinical conditions are classified and relationships between the terms are established and represented. In contrast, the present invention relates to a natural language processing system for extracting *genomics* text data that relates to interactions among genes and proteins followed by computer representation of such information. Applicants assert that such a system relating to genomics text would contain a much larger data set and would be a less structured system than that associated with medical applications, thus, it would not have been obvious that such a system could be successfully adapted from a system relating to medical applications text.

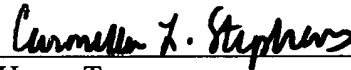
Applicants have amended the claims to encompass a *computerized* method for extracting information on interactions between biological entities from natural-language genomics text data. Thus, with regard to Kim, Kim fails to disclose or suggest such a *computerized* system for representing interactions between genes and proteins derived from genomics text. Therefore, the invention cannot be rendered obvious in view of Kim.

CONCLUSION

Entry of the foregoing amendments and remarks into the file of the above-identified application is respectfully requested. Attached herewith is Appendix A which contains a marked up version of the amendments. The Applicants believe that the invention described and defined by the amended claims is patentable over the rejections of the Examiner. Withdrawal of all rejections and reconsideration of the amended claims is requested. An early allowance is earnestly sought.

Respectfully submitted,

Dated: June 27, 2002



Henry Tang  
Patent Office Reg. No. 29,705

Carmella L. Stephens  
Patent Office Reg. No. 41,328

BAKER BOTTS L.L.P.  
30 Rockefeller Plaza  
New York, New York 10112-4498

Attorney for Applicants  
(212) 408-2539



## APPENDIX A

IN THE TITLE:

[Gene Discovery through Comparisons of Networks of Structural and Functional Relationships Among Known Genes and Proteins] Methods for Extracting Information on Interactions Between Biological Entities From Natural Language Text Data

IN THE SPECIFICATION:

On page 1, please amend the first paragraph as follows:

This application is a continuation-in-part of pending application Serial No. 09/327,938 filed June 8, 1999 which claims priority to provisional patent application Serial No. 60/129,469 filed April 15, 1999. The invention described herein was funded in part by a grant from the National Library of Medicine, namely, Grant Number's LM06274 and LM05627. The United States Government may have certain rights to the invention. The present specification contains a computer program listing which appears as a microfiche Appendix H.

On page 34-35, please amend the last paragraph which continues on page 35 and as follows:

Known motifs/domains for proteins may also be collected using the flat file versions of major protein databases, such as SwissProt [(http://expasy.hcage.ch/sprot)] and the non-redundant database of NCBI [(http://www3.ncbi.nlm.nih.gov)]. The databases can be downloaded and searched for the keywords "motif" and "domain" in the feature tables of proteins. In addition, existing databases of motifs and domains, such as BLOCKS [(http://dupsas.Weizmann.ac.il/bcd/bcdparent//databanksblocks/hfml)] and pfam [(http://www.sanger.ac.uk//software/pfam; http://pfm.wustl.edu)], can be

downloaded (Henikoff et al., 1991, NAR 19:6565-6572). Still further, it is understood that any publically available database containing gene/protein sequences may be utilized to generate the specialized databases for use in the practice of the present invention.

On page 44, please amend the first full paragraph as follows:

To construct a reconciled tree according to the invention, the first step comprises a search for homologs in a publicly or privately available database such as, for example, GenBank, Incyte, binary BLAST databases, Swiss Prot and NCBI databases. Following the identification of homologous sequences a global alignment is performed using, for example, the CLUSTALW program. From the sequence alignment a gene tree is constructed using, for example, the computer program CLUSTLAW which utilizes the neighbor-joining method of Saito and Nei (1997, Mol. Biol. Evol. 4:406-425). Construction of a species tree is then retrieved from, for example, the following [web site: <http://www.3.NCBI.NLM.NIH.GOV/taxomy.tax.html>].

On page 66, please amend the first and second paragraph as follows:

Identification of a putative apoptosis-related human gene began with an identification of all genes in *C. elegans* that contained either a POZ or kelch domain. A subset of these genes is shown in Figure 13. Hidden Markov Models (HMM) for the POZ and Kelch domains were built as follows. Starting with POZ and kelch sequences from the *Drosophila* kelch protein (gi 577275) homologs were identified in other protein sequences using the BLASTP program. The resulting sequences showing significant similarity (e-value less than 0.001) were aligned using CLUSTALW program and the alignments were used to build Hidden Markov Models with HMMER-2 package (Krogh et al., 1995) [, <http://hmmer.wustl.edu/>]. A computer printout listing of HMM

models of tumor suppressors appears as a Microfiche H to the present specification.

[(See, <http://hmmer.wustl.edu>; Chapter 2, which is incorporated by reference herein in its entirety, for a detailed description of HMM models)].

The resulting models were used to search through a database collection of *C.elegans* protein sequences. The domain structures of proteins having either a POZ or kelch domain were identified using existing collections of protein domains [(e.g., see [http://blocks.fhcrc.org/blocks/blocks release.html](http://blocks.fhcrc.org/blocks/blocks%20release.html), <http://coot.embl-heidelberg.de/SMART/>, <http://www.motif.genome.ad.jp/>)].

IN THE CLAIMS:

Please amend the claims to read as follows:

11. (amended) A computerized method for extracting information on interactions between biological entities from natural-language genomics text data, comprising:

- (i) parsing the genomics text data to determine the grammatical structure of the text data ;and
- (ii) regularizing the parsed text data to form structured word terms.

19. (amended) [ The method according to claim 11, further comprising performing error recovery when parsing of the text data is unsuccessful] A computerized method for extracting information on interactions between biological entities from natural-language genomics text data, comprising:

- (i) parsing the genomics text data to determine the grammatical structure of the text data wherein if said parsing of the test data is unsuccessful error recovery is performed; and

- (ii) regularizing the parsed text data to form structured word terms.

21. (amended) [The method according to claim 11, wherein said tagging step comprises providing the structured data component in a Standard Generalized Markup Language (SGML) compatible format] A computerized method for extracting information on interactions between biological entities from natural-language genomics text data, comprising:

- (i) parsing the genomics text data to determine the grammatical structure of the text data ;
- (ii) regularizing the parsed text data to form structured word terms;  
and
- (iii) tagging the text data with a structured data component derived from the structured word terms wherein said tagging step comprises providing the structured data component in a Standard Generalized Markup Language (SGML) compatible format.